

MH  
IL 99/487  
PCT/IL 99/00487  
09 NOVEMBER 1999  
09/786865

REC'D 16 NOV 1999  
WIPO PCT

PA 154430

# THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

September 28, 1999

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: 60/099,702

FILING DATE: September 10, 1998

## PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH RULE 17.1(a) OR (b)



By Authority of the  
COMMISSIONER OF PATENTS AND TRADEMARKS

*L. Edelean*

L. EDELEN  
Certifying Officer

96/01/60  
US PTO

FOLEY & LARDNER  
Suite 500  
3000 K Street, N.W.  
Washington, DC 20007-5109  
(202) 672-5300

JC541 U.S. PTO  
60/009702  
09/18/98

## PROVISIONAL APPLICATION FOR PATENT

Assistant Commissioner for Patents  
Box Provisional Application  
Washington, D. C. 20231

Sir:

This is a request for filing a PROVISIONAL APPLICATION FOR  
PATENT UNDER 37 CFR 1.53(b)(2).

INVENTOR(S) / APPLICANT(S)			
LAST NAME	FIRST NAME	MIDDLE INITIAL	RESIDENCE (City & either State or Country)
WILF	Itzhak		Neve Monoson, Israel
GREENSPAN	Hayit		Kfar Bilu, Israel
MENADEVA	Ovadya		Bat Yam, Israel
CASPI	Yaron		Ness Ziona, Israel
RAHMANI	Amir		Hod Hasharon, Israel

### TITLE OF THE INVENTION

METHOD OF FACE INDEXING FOR EFFICIENT BROWSING AND SEARCHING OF PEOPLE  
IN VIDEO

In connection with this application, the following are enclosed:

16 Pages of Specification (Optional:  Abstract  Claims 17)

16 Sheets of Drawings

— Assignment to: —

— Statement of Small Entity Status

x Other: Return postcard

Attorney Docket No. 023826/0137

The fee has been calculated as shown below. (Small entity fees indicated in parentheses.)

Filing Fee	\$150 (\$75)	\$150.00
Rule 17(k) fee for non-English text	\$130	
Assignment Recording Fee	\$ 40	
TOTAL FEE		\$150.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

No  Yes, the name of the U.S. Government agency and the Government contract number are: .

A check in the amount of the above TOTAL FEE is attached. The Commissioner is hereby authorized to charge any deficiency or credit any overpayment to Deposit Account No. 19-0741.

Respectfully submitted,

  
Bernhard D. Saxe  
Registration No. 28,665

Date: September 10, 1998  
Docket No.: 023826/0137

## Method of face indexing for efficient browsing and searching of people in video

### U.S. Patent Documents

5,164,922	11/1992	Turk, et al
5,664,227	9/1997	Mauldin, et al.
5,708,767	1/1998	Yeo, et al.
5,655,117	8/1997	Goldberg, et al.
5,579,471	11/1996	Barber, et al.
5,715,325	2/1998	Bang et al.
5,764,790	6/1998	Brunelli et al.
5,450,504	9/1995	Calia
5,6642,431	6/1997	Poggio et al.
5,012,522	4/1991	Lambert

### Other Publications

1. Yeung M, Boon-Lock Yeo and Bede L., Extracting Story Units from Long Programs for Video Browsing and Navigation, International Conference on Multimedia Computing and Systems, June 1996.
  2. Yeung, M. and Bede L., Efficient Matching and Clustering of Video Shots, International Conference on Image Processing, October 1995.
  3. Yow, K.C. and Cipolla, R. , Finding Initial Estimates of Human Face Location, In Proceedings 2<sup>nd</sup> Asian Conference on Computer Vision, Vol 3, pp. 514-518, Singapore, 1995.
  4. Jacquin, A. and Eleftheriadis, A., Automatic location and tracking of faces and facial features in video sequences, Proceedings, International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, June 1995.
  5. Peacocke, R.D. and Graf, D.H., An Introduction to Speech and Speaker Recognition, Computer, Vol. 23 (8), pp. 26-33 (1990).
  6. A.L. Higgins, L.G. Bahler and J.E. Porter, Voice Identification using Nearest-Neighbor Distance Measure, 1993 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 375-378, 1993.
  7. Wilcox, L.D. et al, Segmentation of speech using speaker identification, 1994 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 161-164, 1993.
  8. Gish, H. et al, Segregation of speakers for speech recognition and speaker identification, 1991 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 873-876, 1991.
  9. Rabiner, L.R. and Juang, B., Fundamentals of Speech Recognition, Prentice-Hall.

**ABSTRACT**

A Method for efficient browsing and searching of people in video is described. The method consists of a pre-processing stage which automatically parses and indexes the video content based on facial information. A face detection unit indicates a hypothesis for the presence of a face. A face-tracking unit then tracks via the detected face to extract the related video segment. A representative set of facial viewpoints is extracted from the video segment and characteristic facial features are stored. Each newly detected face is matched with a currently existing face database to augment the database at an already existing entry or to introduce a new face entry. The video face index can be displayed, edited, annotated and browsed efficiently, by person. In searching, the face index is accessed to reduce processing time and to increase recognition accuracy. Extracted audio characteristic data is used to validate face matching across video scenes and to augment the face indexing data for future recognition.

**DESCRIPTION OF THE BACKGROUND OF THE INVENTION****1. FIELD OF THE INVENTION**

The present invention relates to video indexing, logging, browsing and searching. The invention focuses on the automatic parsing and indexing of high-level video content, such as structured objects. Moreover, the invention relates between the indexing scheme and the formation of an intelligent index database for recognition tasks in querying applications. More particularly, the invention describes a particular focus on the automatic parsing and indexing of video content based on facial information for efficient browsing and searching of people in video.

**2. DESCRIPTION OF THE RELATED ART**

The amount of video data stored in multimedia libraries grows very rapidly which makes searching a time consuming task. Both time and storage requirements can be reduced by creating a compact representation of the video footage in the form of key-frames, that is a subset of the original video frames which are used as a representation for these original video frames. Prior art focuses on key-frame extraction as basic primitives in the representation of video for browsing and searching applications.

---

A system for video browsing and searching, based on key-frames, is depicted in Fig. 1A. A video image sequence is input from a video feed module 110. The video feed may be a live program or recorded on tape. Analog video is digitized in video digitizer 112. Optionally, the system may receive digital representation such as Motion JPEG or MPEG directly. A user interface console 111 is used to select program and digitization parameters as well as to

control key-frame selection. A key-frame selection module 113 receives the digitized video image sequence. Key-frames can be selected at scene transitions by detecting cuts and gradual transitions such as dissolves. This coarse segmentation into shots can be refined by selecting additional key-frame in a process of tracking changes in the visual appearance of the video along the shot. A feature extraction module 114 processes key-frames as well as non key-frame video data to compute key-frames characteristic data. These data are stored in the key-frames characteristic data store and are accessed by a video search engine 116 in order to answer queries generated by the browser-searcher interface 117. Such queries may relate to content attributes of the video data such as color, texture, motion and others. Key-frame data can also be accessed directly by the browser 117. In browsing, a user may review key-frames instead of the original video, thus reducing storage and bandwidth requirements.

In an edited video program, the editor switches between different scenes. Thus, a certain collection of M video shots may consist only of N<M different scenes such that at least one scene spans more than one shot. Prior art describes how to cluster such shots based on their similarity of appearance for video browsing applications. It has become standard practice to extract features such as color, texture and motion cues, together with a set of distance metrics, and then utilize the distance metrics in the related feature spaces (respectively) for determining the similarity between key-frames of the video contents or shots. In this scenario, the video content is limited to the definition in the low-level feature space. What is missing in the prior art is the automatic extraction of high-level object-related information from the video during the indexing phase, so as to facilitate future searching applications.

In current systems for browsing and automatic searching, which are based on key-frames, the key-frames extraction and the automatic searching are separate processes. Combining the processes in a unified framework means taking into account high-level user-queries (in search mode), during the indexing phase. Spending more effort in a more intelligent indexing process proves beneficial in a short turn around rate in the searching process.

In automatic searching of video data by content, detecting and recognizing faces is of primary importance for many application domains such as news. Prior art describes methods for face detection and recognition in still images and in video image sequences.

A prior art method of face detection and recognition in video is depicted in Fig 1B. A face detection module (122) operates on a set of frames from the input video image sequence. This set of frames may consist of the entire image sequence if the probability of detection is of primary importance. However, the face content in the video does not change every frame. Therefore, to save computational resources a subset of the original sequence may be utilized.

Such a subset may be derived by decimating the sequence in time by a fixed factor or by selecting key-frames by key-frame extraction module 121. The detected faces are sequentially stored in the face detection data store 123.

For each detected face, face features are extracted (124), where the features can be facial feature templates, geometrical constraints, and global facial characteristics, such as eigen-features and other known algorithms in the art. The face representation can be compared to a currently awaiting search query, or to a predefined face database, or alternatively it can be stored in a face feature database (125) for future use. By comparing face characteristic data from the database or from a user-defined query with the face characteristic data extracted from the video, the identity of people in the video can be established. This is done by the face recognition module 126 and recorded in the video face recognition report 127 with the associated confidence factor.

Several algorithms for face recognition are described in the prior art. In particular one prior art embodiment uses a set of geometrical features, such as nose width and length, mouth position and chin shape; Another particular method is based on template matching. One particular embodiment represents the query and the detected faces as a combination of eigen-faces.

In a co-pending application by the same assignee, entitled "A method of automatic extraction of key-frames from a video sequence", a method of key-frame extraction is described. The application discloses a method for post-processing of the key-frame set so as to optimize the set for face recognition. A face detection algorithm is applied to the key-frames and in the case of a possible face detected, the position of that key-frame along the time axis is possibly modified to allow a locally better view of the face. This application does not teach how to link between different views of the same person or how to globally optimize the different views retained of that person.

Figure 1C shows a simple sequence of video scenes and the associated face content, or lack of it. In this example, some people appear in several scenes. Additionally, some scenes have more than one person depicted. Figure 1D depicts the results of a sequential face-indexing scheme such as the one depicted in Figure 1B. Clearly this representation provides a highly redundant description of the face content of the video scenes. In future recognition tasks, each frame will be processed independently.

In a dynamic scene, a person may be visible for only a part or several parts of the scene. During a scene and across the scenes a person generally has many redundant views, but also several different views. In such a situation it is desirable to prune redundant views on the one hand, yet to increase the recognition robustness by comparing the user-defined query against all the available different views of the same person.

Also during a video segment a person may go from a position where he can be detected to a position where he is visible but cannot be detected by automatic processing. In several applications it is useful to report the full segment of visibility for each recognized person.

Prior art does not teach how to detect and index face instances in a video sequence to support these desirable features. In particular, prior art does not teach how to parse an input video stream into face segments and how to link between similar face segments. Furthermore, prior art does not teach how to extract a representative face index and face frame-set, with multiple views for each person, such that face regions can be later detected and recognized with high probability of success.

### SUMMARY OF THE INVENTION

The general problem solved by this invention is that of parsing a video stream into face/no face segments, indexing and logging the facial content, and utilizing the indexed content as an intelligent facial database for future facial content queries of the video data.

The invention introduces the use of a high-level visual module in the indexing of a video stream, specifically, the use of human facial information.

It is an object of the invention to provide an output facial index of the video. Another object of the invention is to provide an output log for the detected faces. A still further object is to provide a facial database that accommodates future video search via facial queries into the video archive.

A further object of the invention is to significantly increase the speed of the recognition. Yet another object of the invention is to improve the probability of recognition.

### BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A describes a prior art method of searching in video by key-frame selection and feature extraction.

Figure 1B describes a prior art method of sequential face indexing and face recognition.

Figure 1C describes a sample sequence of video content.

Figure 1D presents the results of face detection applied to the arrangement of Figure 1C, organized as a sequential index.

Figure 2 describes the browsing and searching system with face indexing, as introduced in this invention.

Figure 3 presents the face index results as pertaining to the example of Figure 1C and as generated by the present invention.

Figure 4 shows a preferred embodiment of a face-track data structure.

Figure 4a shows a set of face characteristic views selected as taught by the present invention from a face track.

Figure 5 describes the process of generating a face track and extracting associated characteristic views and characteristic data.

Figures 6 describes the overall framework for generating tracks of faces from a video image sequence.

Figure 7 describes the processing flow for tracking a single detection result.

Figure 8 describes a scheme of correlation tracking for the eyes.

Figure 9 shows a preferred embodiment for selecting Face\_characteristic\_views.

Figure 10 shows the extraction of audio data corresponding (timewise) to the video data.

Figure 11 shows how to combine the audio track with the face index to create an audio-visual index.

Figure 12 describes the linking and information merging stage as part of the face indexing process.

#### DETAILED DESCRIPTION OF THE INVENTION

It is the purpose of the present invention to teach a method of generating an indexed database of faces that accommodate face (people)-based queries in video search applications.

A system for video browsing and searching of face content (or any other high-level object) is depicted in Fig. 2. A video image sequence is input from a video feed module 210. The video feed may be a live program or recorded on tape. Analog video is digitized in video digitizer 215. Optionally, the system may receive digital representation directly. The video source, the program selection and digitization parameters and face-indexing selection parameters are all controlled by the user from an interface console 230. A subset of the video sequence is input

to an indexing module 220, which is implemented as taught by the present invention. Computed face indexing data are stored in the face indexing data store 250. A graphical representation of the frame indexing data can be displayed and browsed by browser-searcher module 240. Such a graphical representation is depicted in Figure 3. Additionally, the face-indexing data can be edited by a face index editor (270) to correct possible errors that may occur during automatic indexing. Such errors can originate from false face detection that is identifying non-face regions as faces. An additional form of error is an over-segmentation of a certain face: two or more instances of the same face fail to be linked between appearances and thus generate two or more index entries. These errors are handled reasonably by any recognition scheme: false faces will generally not be recognized and over-segmentation will result in somewhat additional processing time and possibly reduced robustness. In the applications in which the generated index will be queried frequently, it is cost-effective to let an editor review the graphical representation of the face index, to delete false alarms and to merge entries originating from the same person.

In a preferred embodiment the editor can annotate the face index by specifying the name of the person or linking between the face instance and another database entity. This embodiment provides a method of semi-automatic annotation of video by first generating a face-index as in Figure 3 and then manually annotating only the index entries. Thus, the annotation becomes immediately linked to all tracks in the video, which correspond to the specific index entry. Since the number of frames where a specific person is visible is much larger than the number of index entries and since the spatial information, that is location within the frame is readily available, a major effort saving is achieved.

Once the face indexing data is stored, a video search engine 260 can access it in order to process face-based queries.

Figure 3 depicts a sample face index that is generated by a particular embodiment of the present invention for the example depicted in Figure 1C.

The face index information, as extracted following face detection and face tracking modules, is stored in a general data structure. Figure 4 shows a preferred embodiment of a face track data structure. Processing a face track consists of tracking the face between detection frames to produce contiguous video segments [Fs,Fc] as well as face track coordinates to be associated with each frame in a video segment. In addition, the face track data structure includes:

- Face Characteristic Views, which are the visually distinct instances of the face in the track.
- Face Frontal Views, which are those Face Characteristic Views, classified as frontal. Frontal faces have better chances of being recognized properly.
- Face Characteristic Data that is attributes computed from the face image data and stored for later comparison with the corresponding attributes extracted from the query image. In a preferred embodiment face characteristic data include eye, nose and mouth templates. In another preferred embodiment face characteristic data include image coordinates of geometric face features. In another preferred embodiment, face characteristic data include coefficients of the eigen-face representation.
- Audio characteristic data that can be associated with the face track.

Figure 4a shows a set of face characteristic views selected as taught by the present invention from a face track. A star denotes face frontal views selected as taught by the present invention.

Figure 5 shows a process of generating a face track and the associated characteristic data from a video image sequence. The processing steps can be initiated at each frame of the video or more likely at a sub-set of the video frames, selected by module 510. The sub-set can be obtained by even sampling or by a key-frame selection process.

In each frame of the subset a face-like region detection method (520) is applied. Preferably, this detection method as taught by prior art, locates facial features. Such features generally consist of eye features and additionally mouth features and nose features.

In Figure 5, the present invention teaches the formation of a face track structure for a single face region. Starting with grouped facial features as output by 520, these features are tracked over time (that is from frame to frame). Preferably, the facial features are tracked from frame to frame (530) by correlation tracking as known in prior art. Both forward and backward tracking is used to extract entire face segment. The result of face tracking is a face track which is a video segment and also the face track coordinates.

Further processing of the video track consists of selection of Face\_Frontal\_Views (540), expanding the selection to a fuller Face\_Characteristic\_VIEWS (550), and extracting Face Characteristic Data (560).

Figures 6, 7 and 8 describe a method of generating tracks of faces from a video image sequence. In Figure 6, a set of results from a face detection module 610 initiates tracking

processes 620 for each of the results. Each tracking process results in a single detection face track, which consists of face location coordinates, and confidence values for a range of frame indices, which include the frame of the original detection. Since a given face may be detected multiple time in a shot, the tracks will overlap and a merging step 630 will follow. The uncertainty in face feature location as resulting from the tracking process is negligible with respect to face size. Therefore, in a preferred embodiment, track merging is implemented on the basis of spatial proximity.

In Figure 7, the processing flow for tracking a single detection result is described. Most face detection methods rely on the detection of facial features such as eyes. These features are a natural choice for face tracking. However, features such as eyes are sensitive to head orientation, blinking, etc. To make the tracking more robust with respect to such disturbances, a head tracking scheme is utilized. The head boundary is estimated from the facial feature data (such as eyes) and an appropriate tracking window is initiated around the head.

In steps 740, 750, 760, 770 the eyes and the head are tracked from the detection frame forward and backward in time until either the end of shot is detected or tracking fails. The individual tracks obtained are merged (780) to a single face track.

In Figure 8 a scheme of correlation tracking for the eyes is described. Initialized by a detection event at frame K the tracking reference frame R is set to K (step 810). In tracking, the location of the tracking points is predicted based on trajectory estimates or set to the previous place where the feature has been detected (step 830). In the case of a feature-pair such as the eyes, a similarity transformation can be derived from the two matched points (step 840). To reduce the apparent difference between the current frame and the reference frame (due to zoom, rotation, etc.) a transformed version R' of the latter is used in the actual correlation matching (step 860). The apparent change between the reference frame and the current frame is repeatedly tested (step 870) and when significant, the reference frame is updated.

Figure 9 shows how to select the set of Face\_Characteristic\_VIEWS from a face track: a video segment, which includes location data for the facial features. Additionally, sets of Face\_Frontal\_VIEWS, which capture the most frontal viewpoint views, are selected as shown also in Figure 4a. These views maximize the probability of recognition in a fixed number of recognition trials. From these sets of face views a set of characterizing features for the face, Face\_Characteristic\_Data are derived.

The set of Face\_Characteristic\_VIEWS is a set of face templates at varying viewpoint angles. Contiguous frames of similar face appearance can be reduced to a single face-frame. Frames

that are different enough in the sense that they can not be reconstructed from existing frames via a similarity transformation are included in the set.

Figure 9 shows a preferred embodiment for selecting the Face\_Characteristic\_VIEWS subject to self-similarity criteria:

The process starts in 905 from the start frame of track both as a reference view and as the only member of the Face\_Characteristic\_VIEWS {C}. Given the currently selected reference frame I, the consecutive frame K is compared against I. In 920 the face motion from I to K is computed from the matched facial features. In 930, the face-like region image in K is compensated for the computed face motion from I to K where said motion is computed from the matched facial features (as extracted from the face track data). The face-like region image in K is compensated for the computed face motion (930). The compensated region is then subtracted from the corresponding face image in I (940) and the difference value is used to decide whether K is a new Characteristic\_Face\_View. In a second embodiment, a frame that contains a Frontal face (taken from the Face\_Frontal\_VIEWS as described below) is selected as an initializing frame and a similar process is performed on frontal faces to obtain all Frontal\_Face\_VIEWS.

---

In an anchorperson scene, the database will contain a limited set of views, as there is limited variability in the face and its viewpoint. In a more dynamic scene, the database will contain a large number of entries per face, encompassing the variety of viewpoints of the person in the scene.

---

Using Face\_Characteristic\_VIEWS captures the variability of face appearance in a relatively small number of views and thus enables the recognition process to be less sensitive to the following (and other) parameters:

- sensitivity to the viewpoint angle of the face;
- sensitivity to facial expressions, including opening vs closing of the mouth;
- sensitivity to blinking;
- sensitivity to external distracts, such as sunglasses

---

The Face\_Characteristic\_VIEWS enables the identification of dominant features of the particular face, including (among others):

- skin-tone coloring;
- eye-color;
- hair shades;

- any special marks (such as birth marks) that are consistent

The *Face\_Frontal\_VIEWS* is a set of the more frontal views of the face. These frames are generally the most-recognizable frames. The selection process is implemented by symmetry-and quality-controlled criteria:

In a preferred embodiment the score is computed from correlation values of eyes and mouth candidate regions with at least one eye and mouth template set, respectively. In another preferred embodiment, the quality index depends also on a face orientation score. In that embodiment said face orientation score is computed from a mirrored correlation value of the two eyes. In yet another embodiment, the face centerline is estimated from mouth and nose location. In that embodiment, the face orientation score is computed from the ratio of distances between the left/right eye to the facial centerline. In yet another embodiment, the face quality index includes also a measure of the occlusion of the face. In that embodiment an approximating ellipse is fitted to the head contour. The ellipse is tested for intersection with the frame boundaries. In yet another embodiment, the ellipse is tested for intersection with other regions.

In a preferred embodiment the process of creating a face track structure includes also the process of computing face characteristic data. Prior art describes a variety of face recognition methods, some based on correlation techniques between input templates, and others utilize distance metrics between feature sets. In order to accommodate the recognition process, a set of face characteristic data is extracted from the *Face\_Characteristic\_VIEWS*.

In a preferred embodiment *Face\_Characteristic\_Data* include:

- 
- $F_g$  = Global information; consists of face templates at selected viewpoints, as well as facial component templates, in one implementation of eyes, nose and mouth.
  - $F_f$  = Facial feature geometrical information indicative of the relationships between the facial components;
  - $F_u$  = Unique characteristics, such as eyeglasses, beard, baldness, hair color;
  - $F_a$  = Audio characteristic data.
- 

Many video sequences include multiple faces. For the case of more than one face candidate detected in frame T, a locality check is pursued to check that the candidates are sufficiently distant and do, in fact, represent separate faces. Each face-like region is then tracked, a face-segment is extracted and a face-frame set is selected, as described above.

Following the definition of a face segment, audio information in the form of the Audio characteristic data is incorporated as additional informative characteristic for the segment,  $F_a$ . It is a purpose of the present invention to associate audio characteristic data with a face track or part of a face track. By combining the results of visual-based face recognition and audio-based speaker identification, the overall recognition accuracy can be improved.

Figure 10 shows a timeline and video and audio data, which correspond, to that timeline. The face/no-face segmentation of the video stream serves as a master temporal segmentation that is applied to the audio stream. The audio segments derived can be used to enhance the recognition capability of any person recognition system built on top of the indexing system, which is taught in the present invention.

It is a further purpose of the present invention to match audio characteristic data, which correspond to two different face tracks in order to confirm the identity of the face tracks. Therefore, the present invention utilizes prior art methods of audio characterization and speaker segmentation. The latter is required for the case the audio may correspond to at least two speakers.

Figure 11 shows how to combine the audio track with the face index to create an audio-visual index. The present invention uses prior art method in speech processing and speaker identification.

Prior art models speakers using a network of hidden Markov models (HMM). Each speaker is modeled using a HMM consisting of states corresponding to the acoustic patterns produced by the speaker. In addition to modeling speakers, HMMs can also be used to model silence and non-speech signals such as laughter.

In a preferred embodiment no-prior knowledge of the speakers is assumed. In that embodiment, unsupervised speaker segmentation is done using an iterative algorithm. Parameters for the speaker HMMs are first initialized using a clustering procedure and then iteratively improved using the Viterbi algorithm to compute segmentation.

In a preferred embodiment, the segmentation of the audio track is aided by visual cues from the face indexing process 1130. In particular, the audio is partitioned with respect to the audio content (1130). For example when an entire shot includes a single face track, the initial hypothesis can be a single speaker. Once verified, the audio characteristic data (such as the

HMMs parameters) are associated with that face. In another example, when an entire shot includes only two faces, the initial hypothesis can be two speakers. Once verified, the audio characteristic data of a speech segment are associated with the face of highest mouth activity as computed by the visual mouth activity detector 1120. In yet another embodiment, audio characteristic data are matched against mouth movement.

The present invention teaches how to incorporate the extracted information from a face segment into a face database, providing links between similar face segments and merging the information. The linking and information-merging stage, as part of the face indexing process, is depicted in Figure. 12. If the face database is empty (1210), the detected face segment initializes the database, providing its first entry. Otherwise, distances are calculated between the new face segment characteristics and each face entry in the database (1220). An overall distance measure is calculated as a function of the individual distance components, in one embodiment being the weighted sum of the distances. Distance measures are ranked in increasing order (1230). The smallest distance is compared to a *Similarity threshold* parameter (1240) to categorize the entry as a new face to be entered to the database, or as an already existing face, in which case the information is merged to an existing entry in the database.

The embodiment in the present invention has been restricted to the indexing of facial content in video sequence. However, the methods taught by the invention can be readily modified to include indexing for browsing and searching of other structured objects. As long as a correspondence can be established for an object by a combination of tracking and matching across different video shots, a similar index structure, which consists of Object log, Object characteristic views and Object Characteristic Data can be constructed. Furthermore, this tracking and matching steps can be used to automate the insertion of a new object into the database.

As a simple example consider matching man-made objects such as building. Suppose that we want to recognize the White House in a video program on the president. The program will include several video scenes in which the White House is visible. Each such scene includes at least one object track of the White House. Since feature points such as corners characterize man-made objects, such feature points can be used to track the objects from frame-to-frame. Additionally, by matching sets of feature-points across video scene, correspondence between views can be established across video shot boundaries. Given tracks of objects, Object\_Characteristic\_View are selected and Object\_Characteristic\_Data are computed as taught by the present invention. Once the object index has been constructed it can be used to browse and search similar objects in the index rather than in the raw video.

We claim:

1. A method for creating a visual index of persons from a video image sequence comprising of:  
Detecting at least one face in a video frame, and  
Testing at least one face video frame for depicting the same person in at least another face instance in said video image sequence;
2. A method for creating a visual index of persons from a video image sequence comprising of:  
Detecting at least one face in a video frame, and  
Tracking said detected face in said image sequence;
3. A method as in 1 or in 2 where said index includes video frame number data for at least one face in the index.
4. A method as in 1 or in 2 where said index includes video frame location data for at least one face in the index.
5. A method of selecting face characteristic views from a video image sequence comprising of:  
Creating an index of faces from a video image sequence, and  
Selecting visually distinct video frames from the index, which depict at least one person.
6. A method of selecting face frontal views from a video image sequence comprising of:  
Creating an index of faces from a video image sequence, and  
Selecting visually distinct video frames from the index, which depict at least one person such that the face pose in each of these frames is frontal.
7. A method of recognizing faces in video comprising of:  
Creating an index of faces from a video image sequence, and  
Matching a set of at least one query faces against said index.
8. A method as in 5 and 6 where said matching is done against said face characteristic views.
9. A method as in 5 where said visually distinct video frames differ at least by head orientation.
10. A method as in 5 or in 6 where said visually distinct video frames differ at least by mouth appearance.
11. A method as in 5 or in 6 where said visually distinct video frames differ at least by eyes appearance.

---

12. A method of annotating face content in a video image sequence comprising of:  
Creating an index of faces from a video image sequence, and  
Attaching annotation data to the index.
13. A method for creating an audio-visual index of persons from a video image sequence comprising of:

Creating an index of faces from a video image sequence, and  
Matching audio characteristic data to said index.

14. A method as in 13 where said matching comprises of comparing speech time segments with face instance time segment.
15. A method as in 14 where said matching includes also detecting mouth activity for at least one face instance.
16. A method as in 14 where said matching includes also matching speech movements to mouth movement for at least one face instance.
17. A method as in 1 where testing for depicting the same person comprises also of comparing audio characteristic data.

United States Patent & Trademark Office  
Office of Initial Patent Examination -- Scanning Division



Application deficiencies found during scanning:

1. Application papers are not suitable for scanning and are not in compliance with 37 CFR 1.52 because:
  - All sheets must be either A4 (21 cm x 29.7 cm) or 8-1/2" x 11".  
Pages \_\_\_\_\_ do not meet these requirements.
  - Papers are not  flexible,  strong,  smooth,  non-shiny,  durable, and  white.
  - Papers are not  typewritten or mechanically printed  in permanent ink  on one side.
  - Papers contain improper margins. Each sheet must have a left margin of at least 2.5 cm (1") and top, bottom and right margins of at least 2.0 cm (3/4").
  - Papers contain hand lettering.
2. Drawings are not in compliance and were not scanned because:
  - The drawings or copy of drawings are not suitable for electronic reproduction.
  - All drawings sheets are not either A4 (21 cm x 29.7 cm) or 8-1/2" x 11".
  - Each sheet must include a top and left margin of at least 2.5 cm (1"), a right margin of at least 1.5 cm (9/16") and a bottom margin of at least 1.0 cm (3/8").
3. Page(s) \_\_\_\_\_ are not of sufficient  clarity,  contrast and  quality for electronic reproduction.
4. Page(s) \_\_\_\_\_ are missing.
5. OTHER: No Declaration

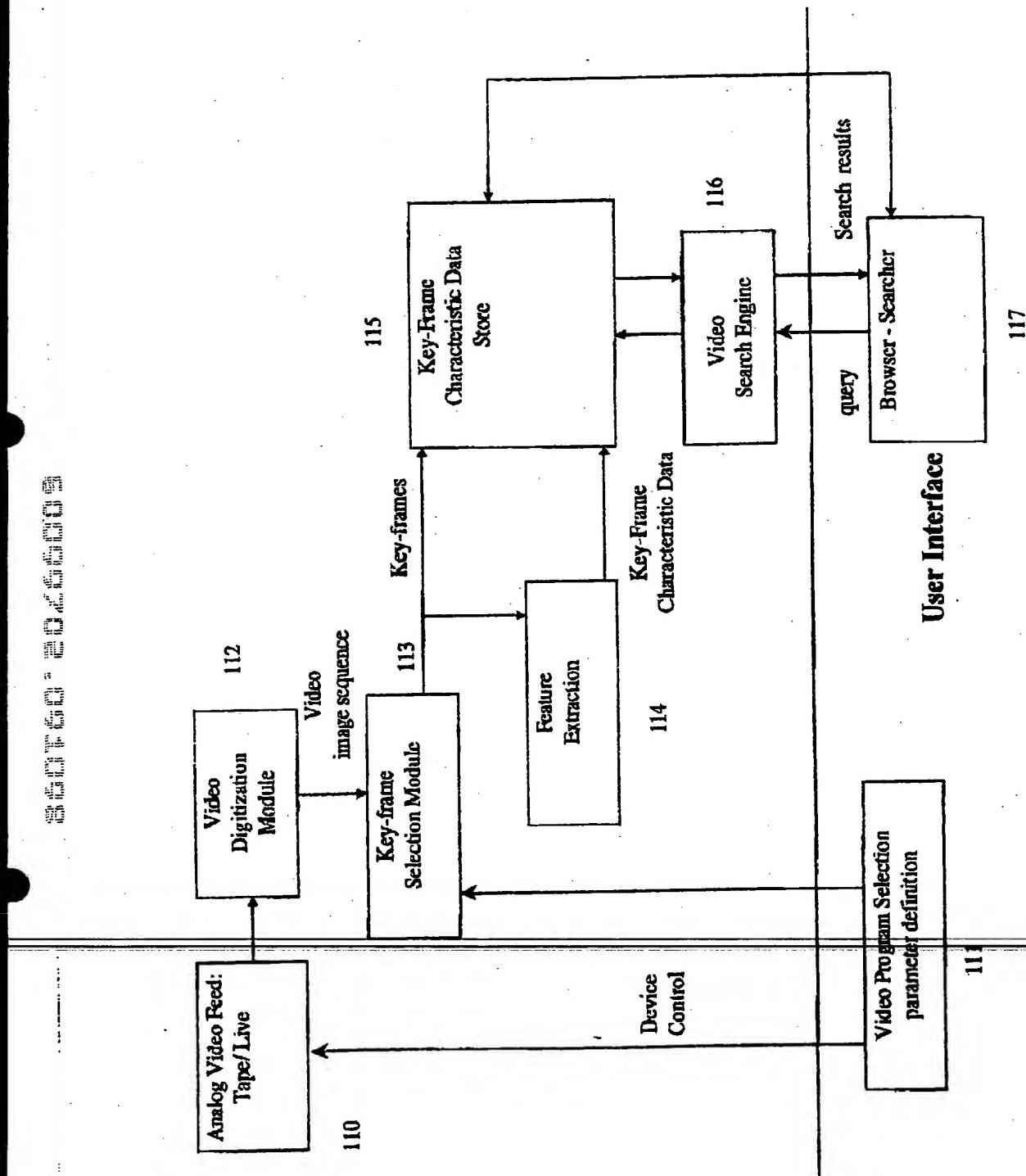
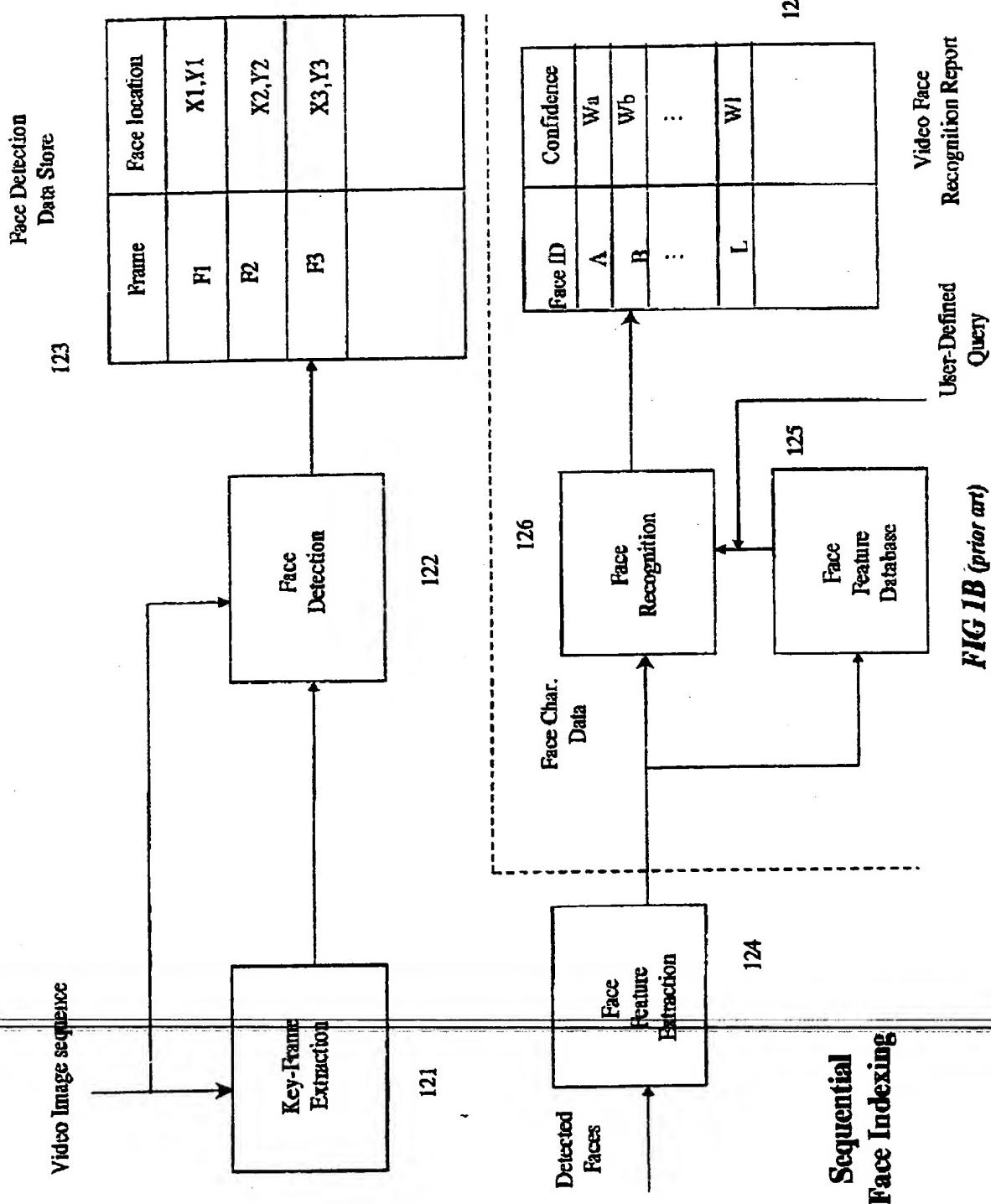
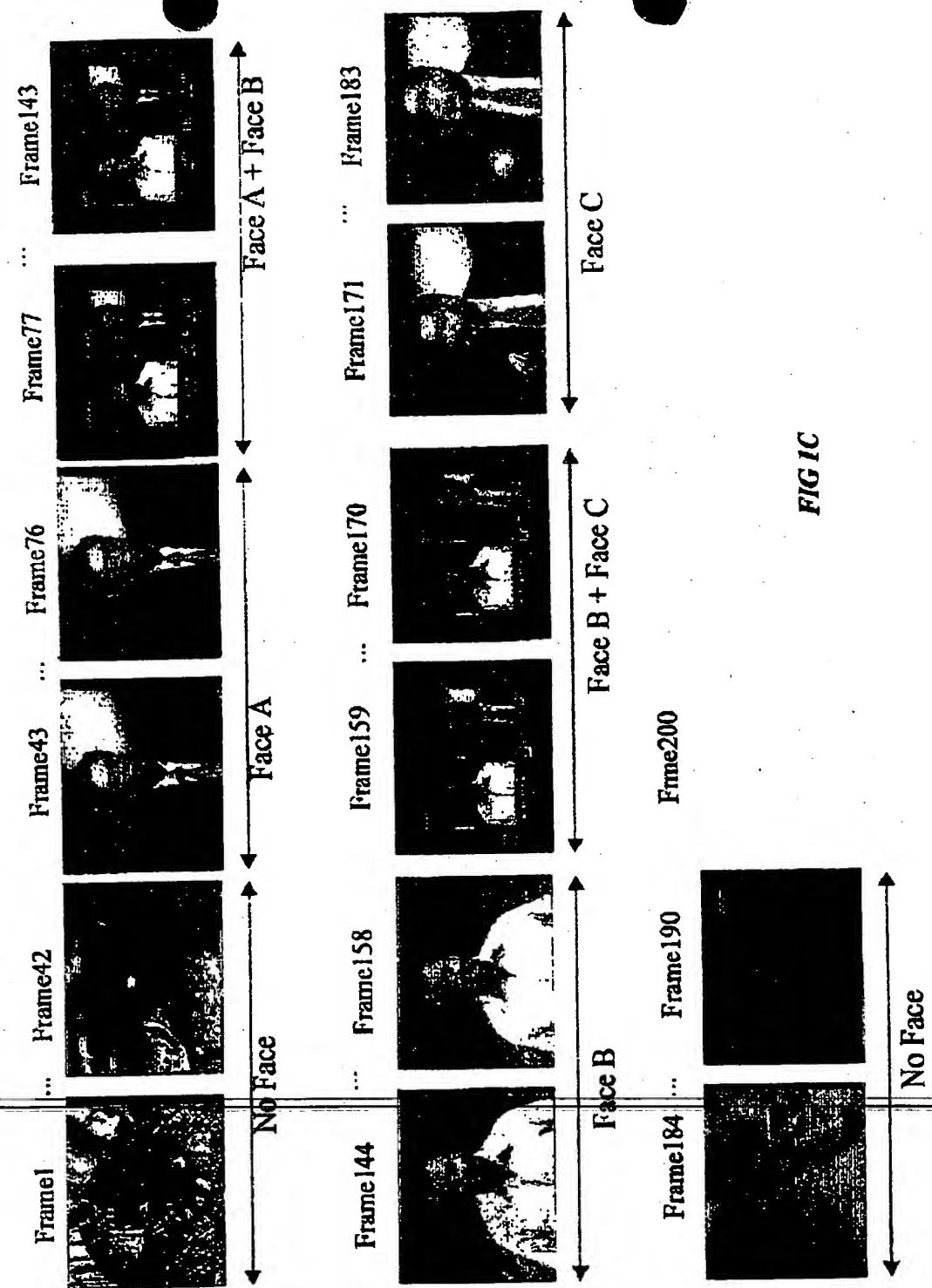


FIG 1A (prior art)





**FIG 1D(prior art)**

<b>Face Instance</b>	<b>Frame Index</b>	<b>Bounding Rectangle</b>
I_1	Frame 43	R_1
...	...	...
I_33	Frame 76	R_k
I_34	Frame 77	R_k+1
I_35	Frame 77	...
I_36	Frame 78	R_n
...	...	R_n+1
I_198	Frame 143	R_s
...	...	...

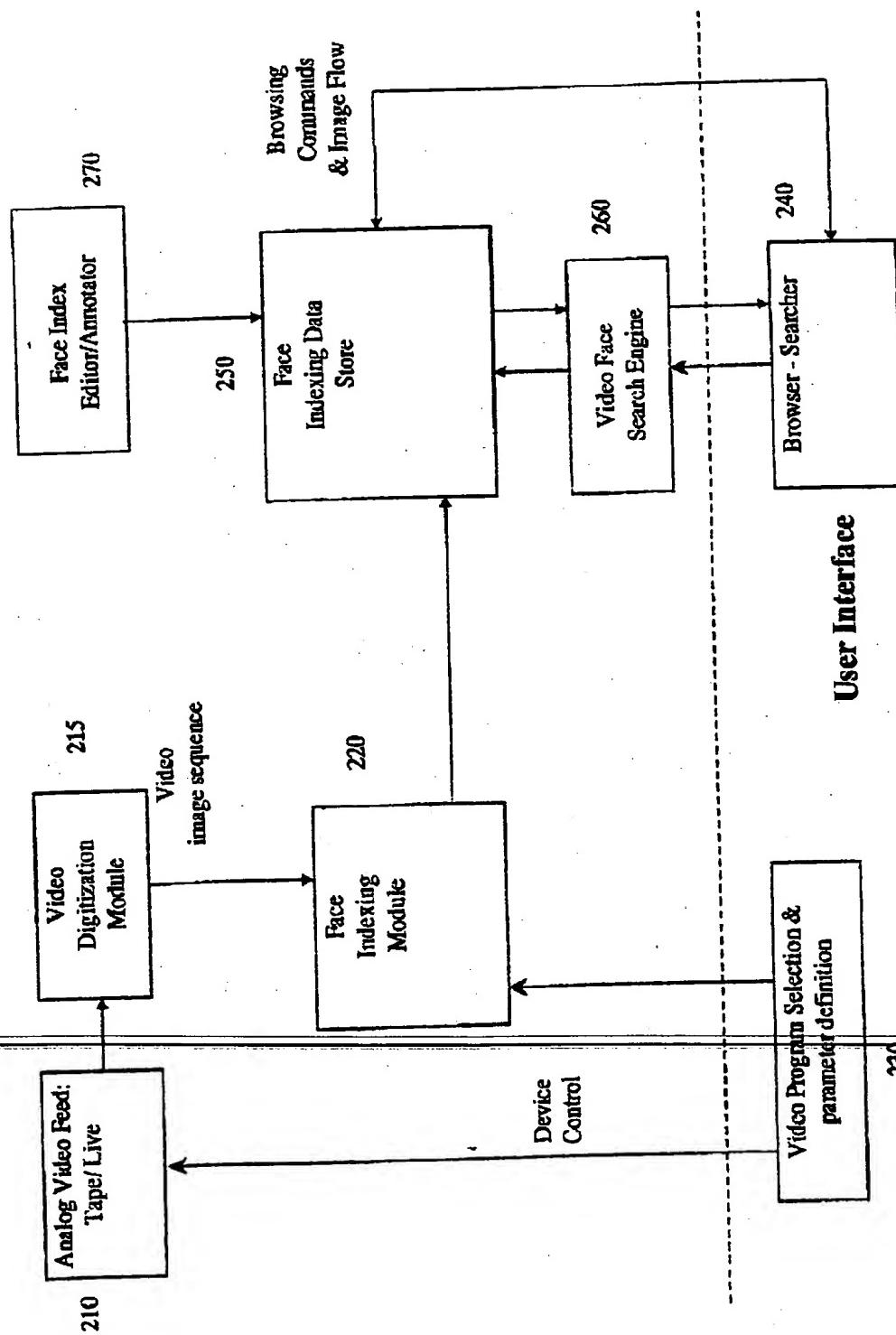


FIG 2

Face Label	Face Track	Frame Index	Bounding Rectangle Per face instance
A	A1	43 - 76	Per face instance
	A2	77-143	...
B	B1	77-143	R_k
	B2	144-158	R_{k+1}
C	C1	159-170	...
	C2	171-183	R_n
...	...	...	...

FIG 3

S.SEP.1998 15:58

EITAN, PEARL, LATZER & COHEN-ZEDEK

00405

P.26

SHAFDAN 202600

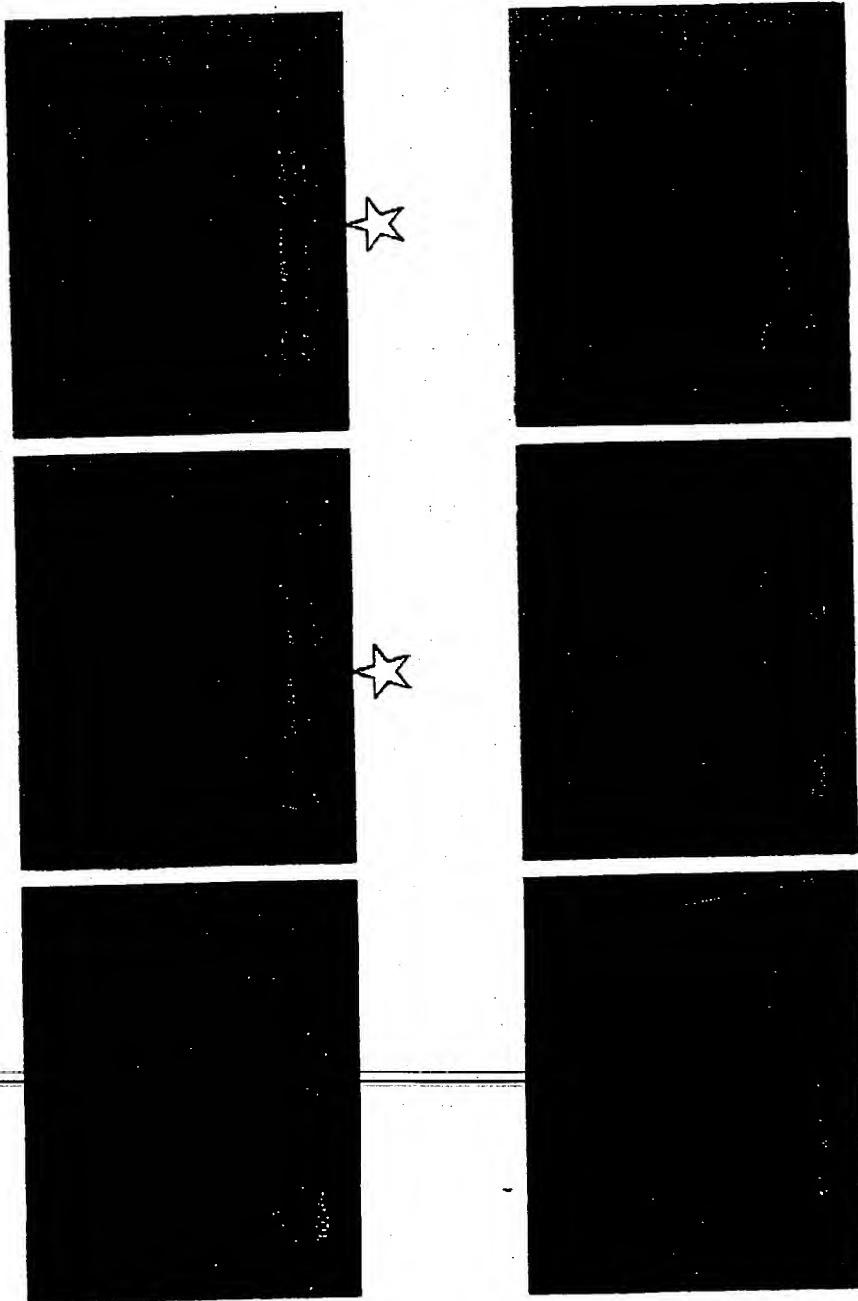


FIG 4a

Start Frame	T <sub>s</sub>
End Frame	T <sub>e</sub>
Face Track Coordinates	
	$\bar{X}_{F_S}, \dots, \bar{X}_{F_e}$
Face_Characteristic_VIEWS	
	Face_Frontal_VIEWS
Face_Characteristic_Data	
	Audio_Characteristic_Data

Face Track Structure

FIG 4

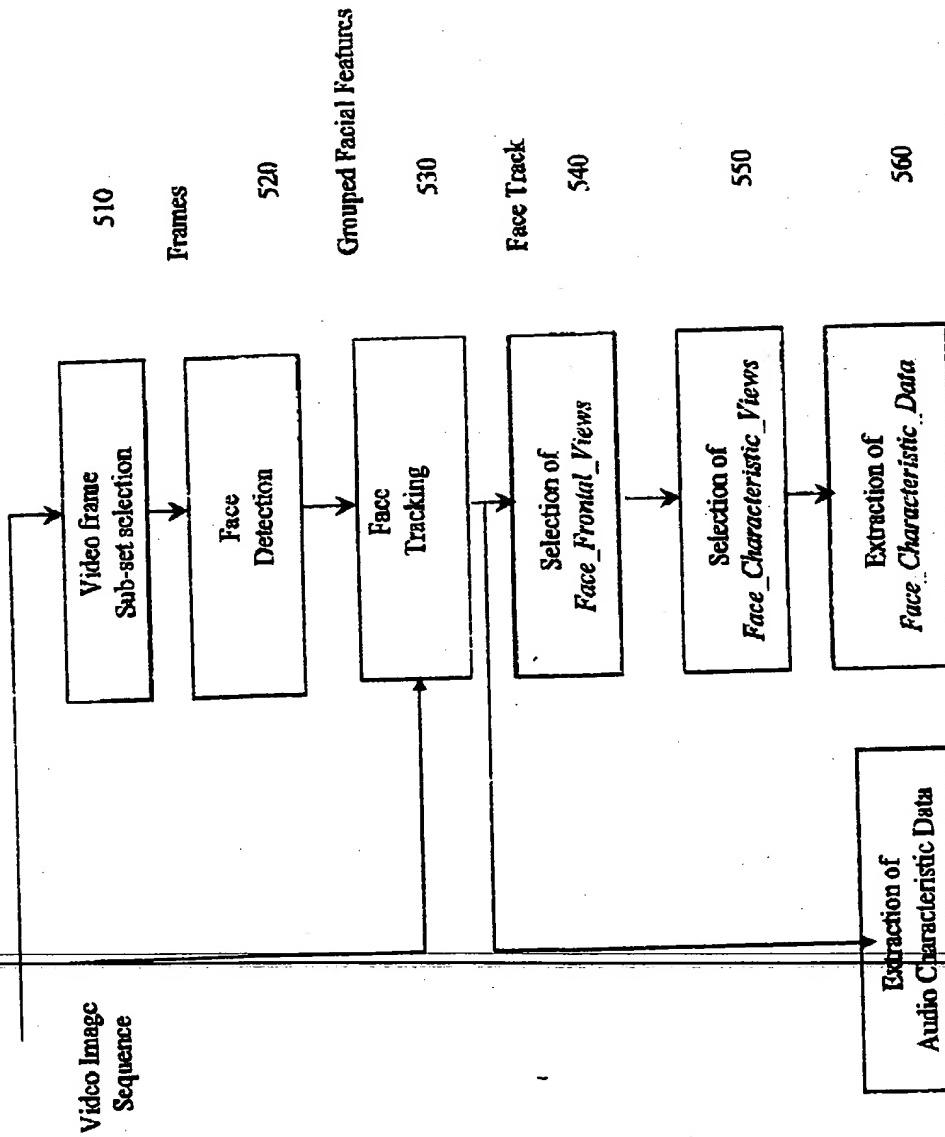


FIG 5

00000000000000000000000000000000

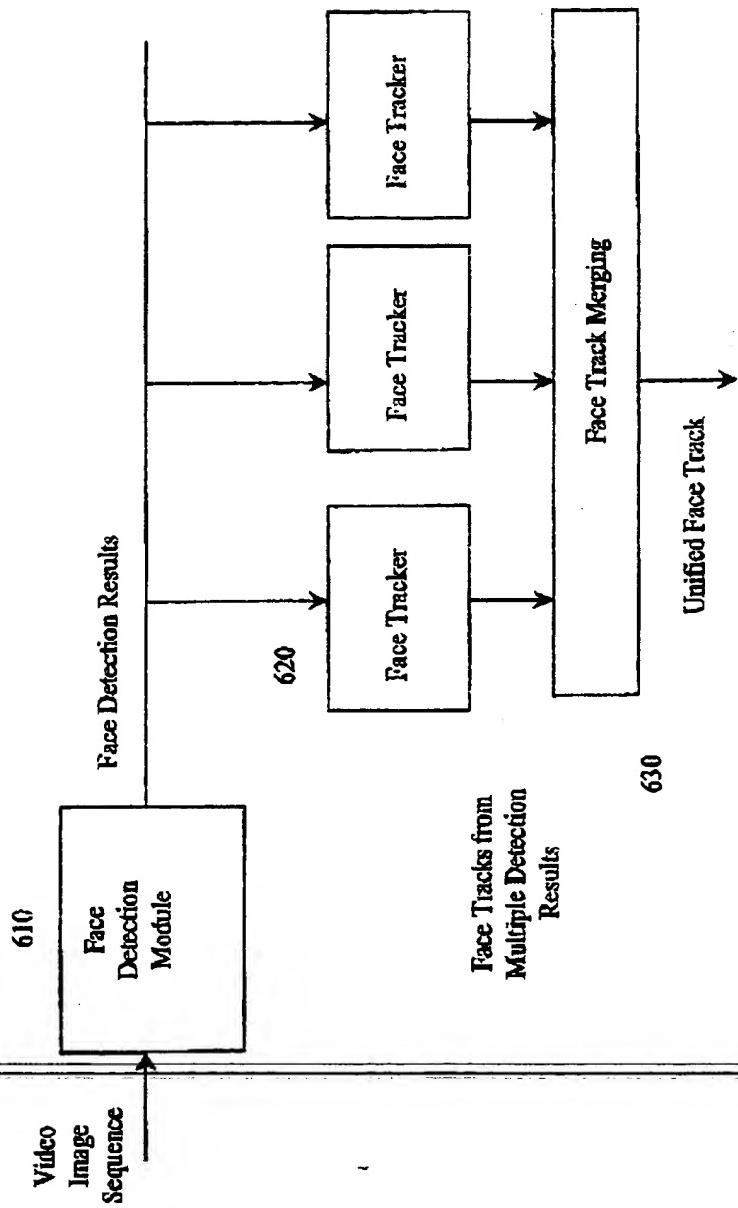


FIG 6

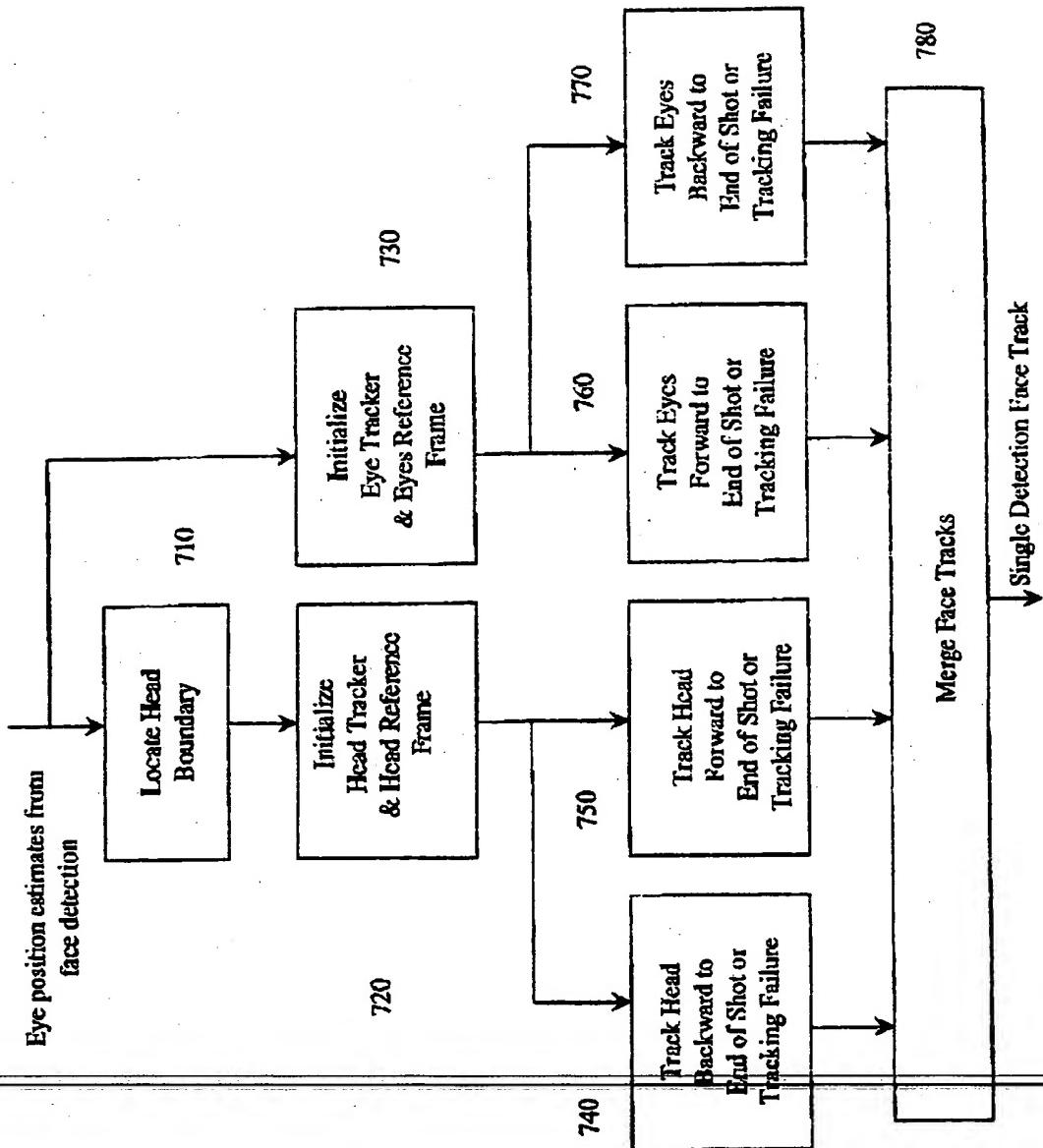


FIG 7

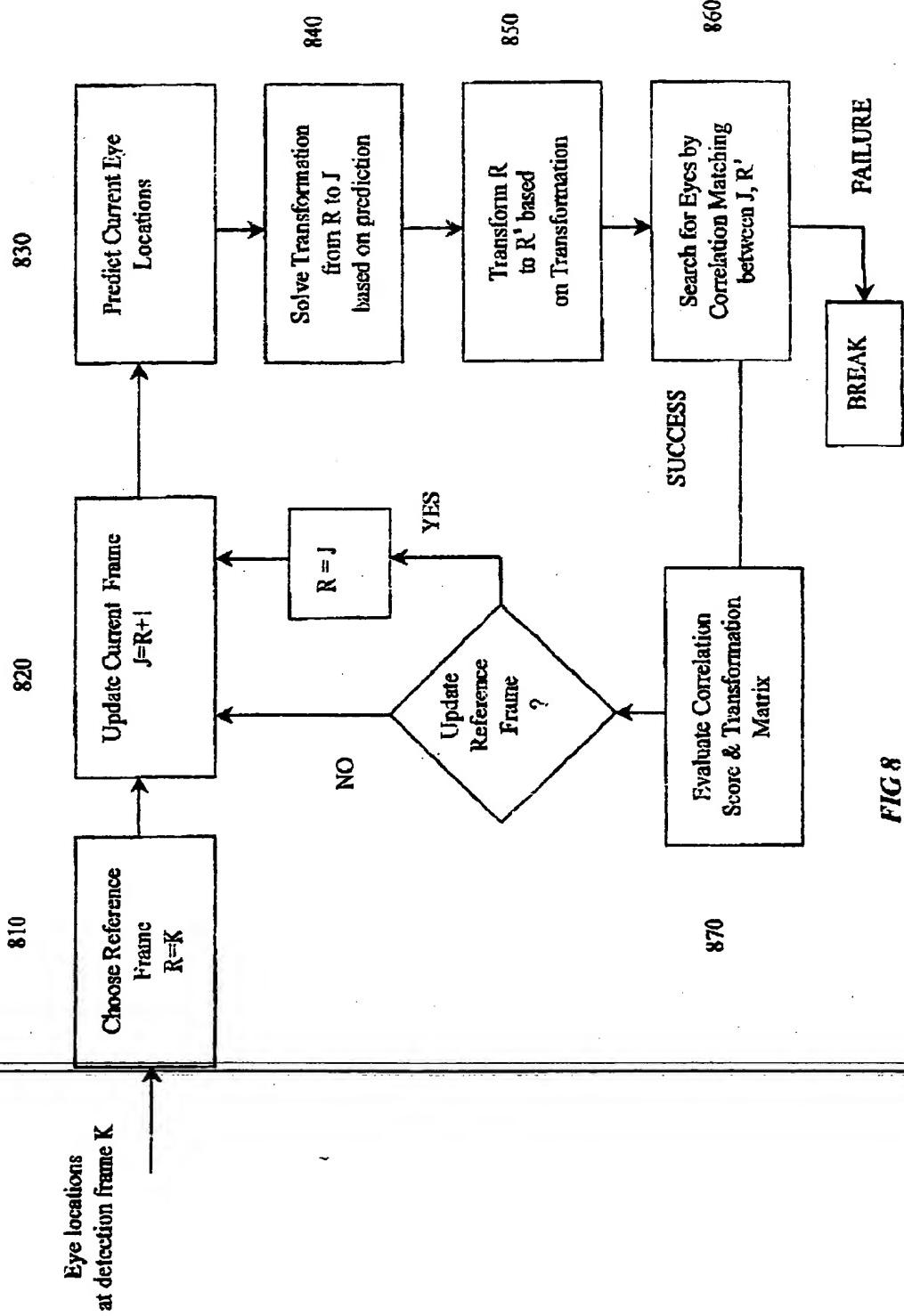


FIG 8

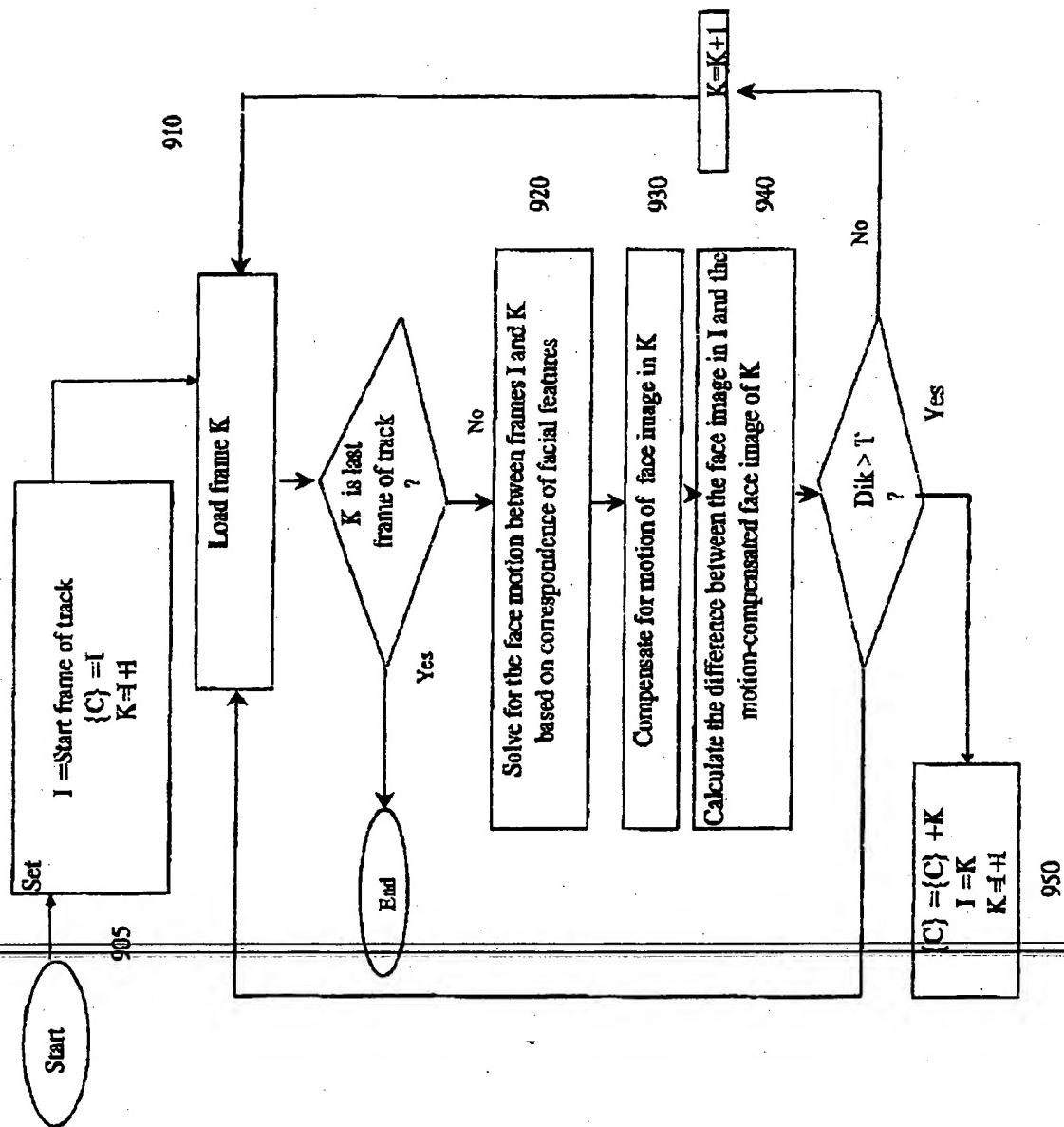


FIG 9

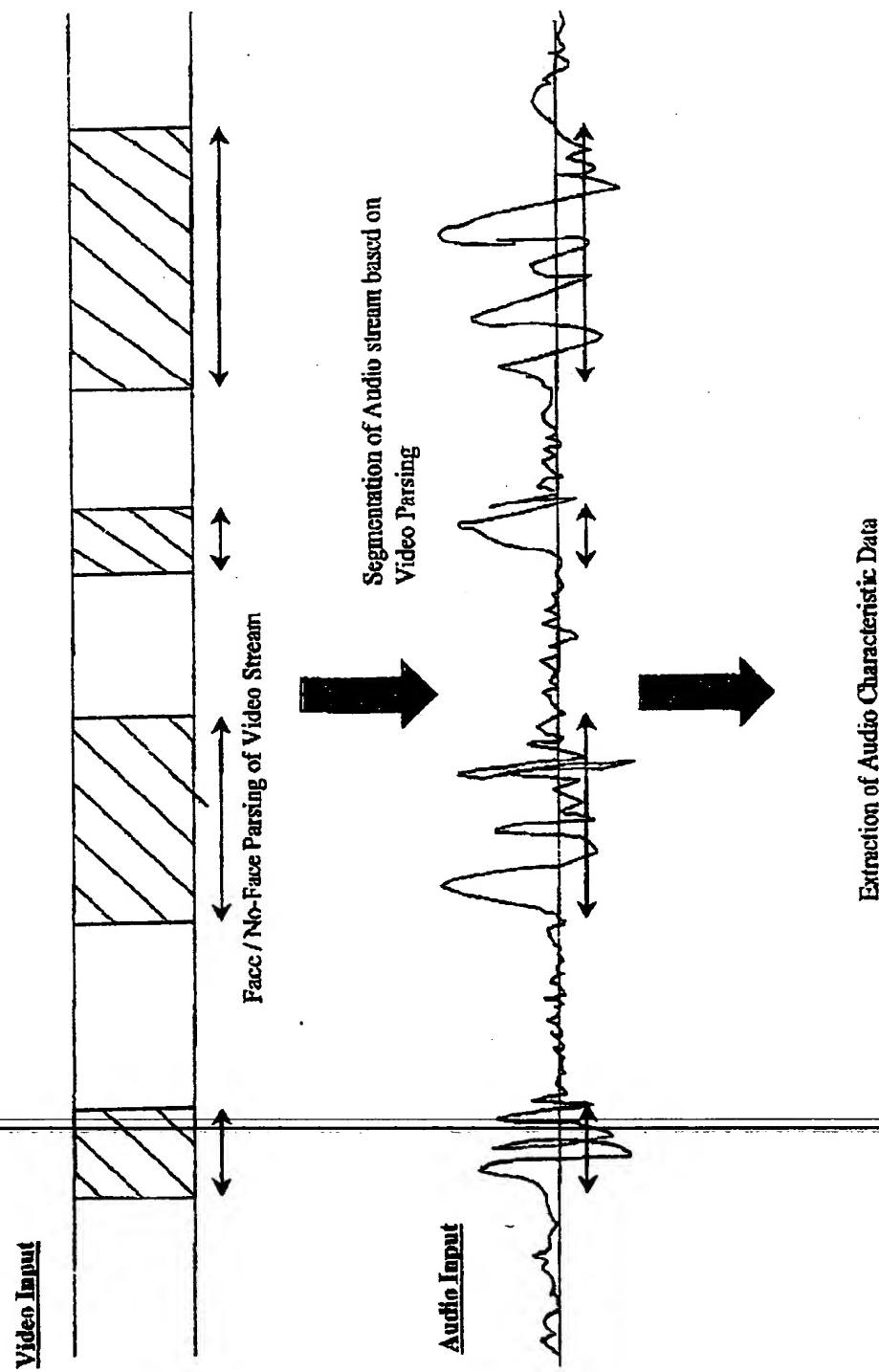


FIG 10

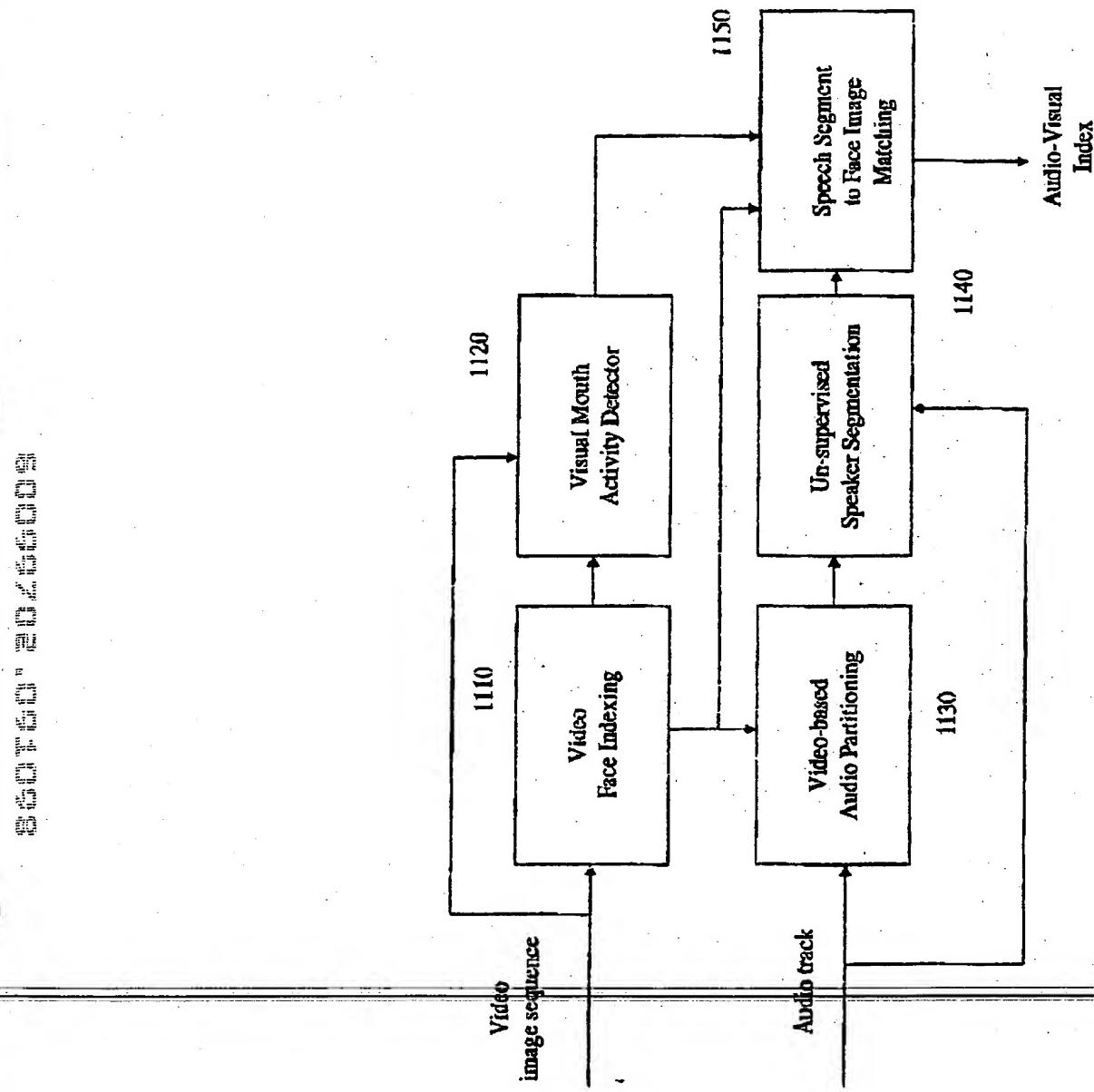


FIG II

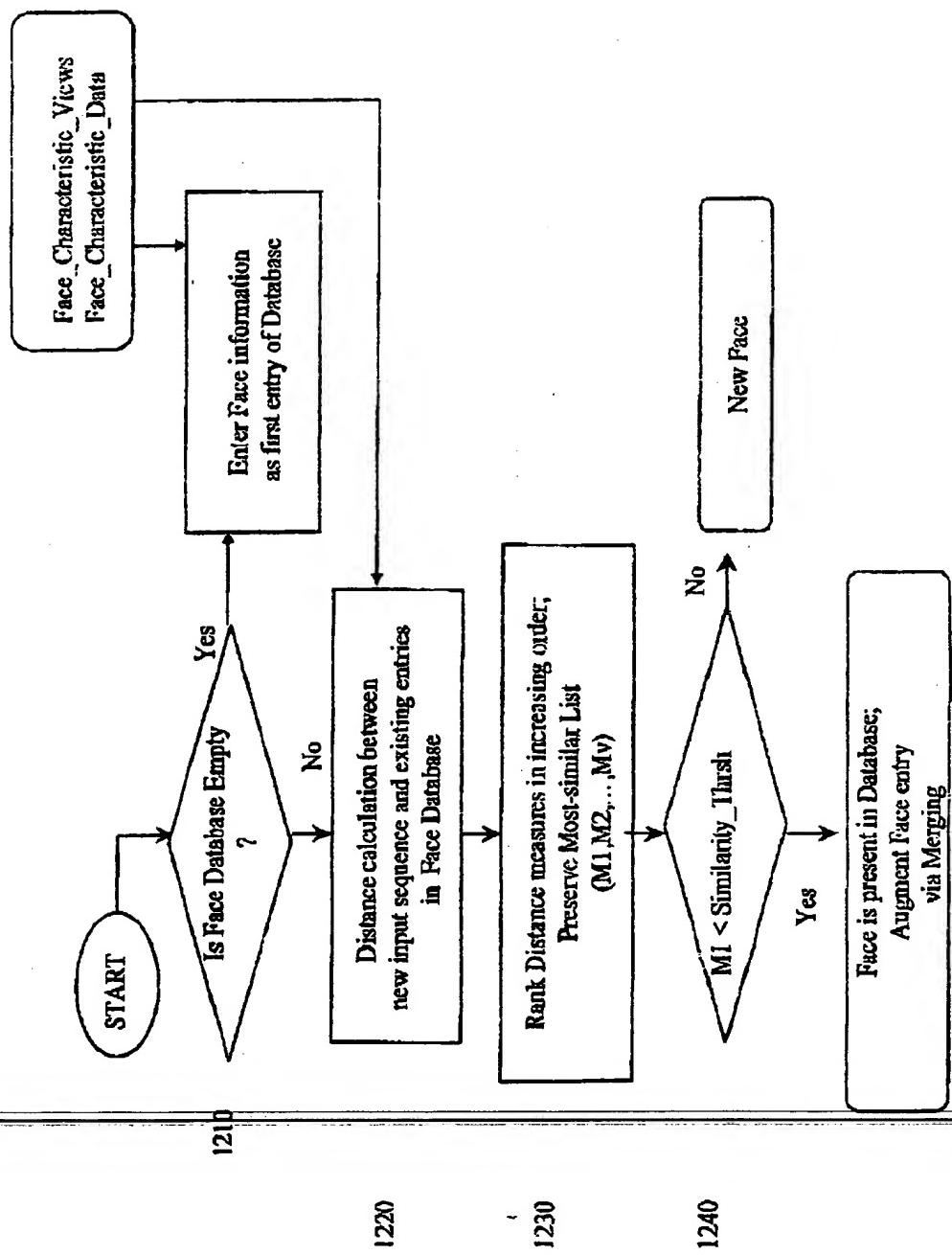


FIG 12